In pursuit of causality in leadership training research: A review and pragmatic recommendations

Robin Martin, University of Manchester, UK

David J. Hughes, University of Manchester, UK,

Olga Epitropaki, Durham University, UK

Geoff Thomas, University of Manchester, UK

Please cite as: Martin, R., Hughes, D.J., Epitropaki, O. & Thomas, G. (in press). In pursuit of causality in leadership training research: A review and pragmatic recommendations. *The Leadership Quarterly.* https://doi.org/10.1016/j.leaqua.2019.101375

Abstract

Although research shows a reliable association between leadership training and positive organizational outcomes, a range of research design issues mean we do not know to what degree the former causes the later. Accordingly, the paper has two main aims. First, to describe the conditions necessary to determine causality in leadership training research and the ability of different research designs to achieve this. Six important, but often ignored, issues associated with determining causality are described (control conditions, sample representation, condition randomization, condition independence, temporal design, and author involvement). Second, to review the extent to which the leadership training literature is able to demonstrate causality. The review shows that the majority of studies do not meet many of the criteria, even the most basic criteria, required to establish causality. Finally, we provide guidelines for designing future research to improve causal identification and is capable of generating meaningful theory and policy recommendations.

"Shallow men believe in luck. Strong men believe in cause and effect" (Ralph Waldo Emerson)

1. Introduction

Leadership training initiatives are highly popular due to the belief that leadership training can lead to (or more precisely *cause*) positive changes in managers' behaviors and ultimately their effectiveness. Given the popularity of leadership training, it is imperative to assess whether, when, and how such training causes change in outcomes in order to advance theory and inform training policy in organizations. As a result, research examining leadership training and development is common (Day, 2001; Day & Dragoni, 2015; Day, Fleenor, Atwater, Sturm, & McKee, 2014; Feldman & Lankau, 2005; Riggio, 2008) and many studies report positive associations between leadership training and a range of work-related outcomes (Aguinis & Kraiger, 2009; Avolio, Reichard, Hannah, Walumbwa, & Chan, 2009; Lacerenza, Reyes, Marlow, Joseph, & Salas, 2017). Further, when effective, leadership development interventions can provide considerable return on investment (Avolio, Avery, & Quisenberry, 2010). However, due to a range of limitations in study design, leadership training studies vary considerably in the extent that they are able to evidence specific causal effects (Podsakoff & Podsakoff, 2019).

Although discussion of causality is common to all areas of research, the leadership training context introduces specific issues that need careful consideration. Undoubtedly, the 'gold standard' for determining causality is the classic experimental design (i.e., participants randomly allocated into treatment and control conditions; Lonati, Quiroga, Zehnder, & Antonakis, 2018). Given an adequate sample size and provided randomization is implemented correctly (i.e., random allocation of leaders to training and control), experimental studies generate two (or more) conditions that are probabilistically similar, meaning that potential confounding variables are evenly distributed and therefore cannot account for differences in outcomes between conditions. Alas, outside laboratory settings (e.g., field and quasi-

experimental studies) researchers seldom have the opportunity to randomly allocate managers drawn from random samples or to control other environmental features because organizations are complex social systems with many processes occurring simultaneously (Shaver, 2019). Managers engage in many activities other than leadership training and the outcomes of interest (e.g., job satisfaction and performance) have multiple determinants. Thus, it is difficult to isolate specific causal factors and identify the proportion of variance in outcomes that is specifically due to, or caused by, the leadership training.

At this point, we should acknowledge that conducting leadership training studies in an applied context is very difficult due to the reliance upon a client organization. Therefore, we acknowledge that leadership training researchers need to balance the need for rigorous research design with contextual constraints, over which they typically have little or no control. Thus, demanding that all leadership training studies reach the highest standards of experimental design is likely to prove counterproductive. Indeed, we believe that all well-designed research that takes into account the ability of the design to determine causality, and tempers conclusions accordingly, can make significant contributions to the 'cumulative body of research' (Shaver, 2019). Nevertheless, as Stouffer (1950) argued "… it is essential that we always keep in mind the model of a controlled experiment, even if in practice we may have to deviate from an ideal model" (p. 356).

Previous reviews tend to focus primarily on training outcomes or the theoretical leadership models underlying training design. Typically, when reporting findings past reviews have tended to aggregate results from studies with varied research designs (e.g., Avolio, et al., 2009; Lacerenza et al., 2017) or make general distinctions between experimental and quasi-experimental designs (e.g., Podsakoff & Podsakoff, 2019). However, if it is the case that leadership training studies vary in terms of their ability to establish causality, then such aggregated analyses can be problematic and produce misspecified estimates that are unable to identify which aspects of training carry the causal effects. In contrast, this paper focuses, for the

first time, on the fundamental research design and methodological features needed to establish the causal effects of leadership training. No matter the design of the leadership training content, if one is unable to establish that the training caused changes in outcomes, then the value of the results is considerably diminished. Therefore, establishing conditions to determine causality is at the heart of the leadership development literature, because, as Aristotle says, "we do not know a truth without knowing its cause". In other words, how can we be confident that a leadership training programme caused positive (or negative) work outcomes if we are unable to evaluate studies (and inform future ones) in terms of their ability to provide causal evidence?

The review makes three important contributions to the leadership training literature. First, we extend the literature by discussing six critical conditions necessary for establishing causal effects and evaluate the extent to which commonly applied leadership training research designs satisfy these criteria. Second, for the first time, guided by these criteria we conduct a state-of-the-science review of the leadership training literature and find that the majority of studies fail to meet the standards necessary for establishing causality. Based on these findings, we identify a number of concerns about the methodological rigor of the literature that have been overlooked by previous reviews. Finally, based on the above, we provide the most comprehensive recommendations to date to guide future leadership training research in terms of establishing causality.

2. Efficacy of leadership training interventions

Leadership training is one type of intervention, under the general umbrella of leadership development activities and interventions, designed to enhance both individual and collective leadership capacity "…to engage effectively in leadership roles and processes" (Day, 2000, p. 582). Day and Dragoni (2015) distinguished between "… *leader development* as the expansion of the capacity of individuals to be effective in leadership roles and processes" and "… *leadership development* as the growth of a collective's capacity to produce direction, alignment, and commitment" (Day & Dragoni, 2015, p. 134). Leader development focuses on individual-

level concepts such as self-awareness, self-efficacy and KSAs, whereas leadership development adopts a more integrated approach focusing on the dynamic interplay between leaders and followers, dyadic relationships, and leadership behavioral constructs such as transformational leadership (e.g., Barling, Weber, & Kelloway, 1996; Riggio, 2008). This distinction has many implications for the content of the training and development process and related outcomes. In this paper, we focus on studies that evaluate leadership training, i.e., when an organization deliberately exposes managers to leadership training activities with the expectation that this should improve leadership capacity and consequently have an impact on relevant work-related outcomes.

Leadership training varies considerably in terms of the mode of delivery (e.g., classroom, coaching), process (e.g., multi-source feedback, reflection, role-playing), and activities (e.g., mentoring, coaching, job assignments, see Day, 2001, for a review). The combination of one or more leadership training activities can form an 'intervention' in a research study sense. More precisely we define leadership training as "… those studies where the researcher overtly manipulated leadership as an independent variable through training, assignment, scenario or other means" (Avolio et al., 2009, p. 764). In research design terms, leadership training is the independent variable and the outcomes of the training are the dependent variable. A range of different types of outcomes can be obtained (e.g., from leaders, followers, HR, clients etc.) either before (pre-test) and/or after (post-test) the training.

A number of meta-analytic reviews show a moderate relation between leadership training and a range of important work outcomes (Avolio et al., 2009; Burke & Day, 1986; Collins & Holton, 2004; Lacerenza et al., 2017; Powell & Yalcin, 2010; Russon & Reinelt, 2004; Taylor, Russ-Eft, & Taylor, 2009). However, there is considerable variation in findings as pointed out by Collins and Holton (2004) "…the effectiveness of managerial leadership development programs varied widely; some programs were tremendously effective, and others failed miserably" (p. 232). In addition, due to their different research questions, these meta-

analyses employ different inclusion criteria and therefore vary considerably in terms of the literature that they cover. They also tend to take a broad conceptualization of leadership development, including studies that cover general managerial skills (such as problem-solving, job design, Collins & Holton, 2004) that might not be germane to leadership. The most recent meta-analysis by Lacerenza et al. (2017) examined 335 independent studies, three times more than the next highest meta-analysis. Overall, the relationship between leadership training and employee outcomes was high on all four of Kirkpatick's (1959) criteria for evaluating training (i.e., reactions, learning, transfer, and results; overall ρ = .76) and this led the authors to conclude that "… leadership training is substantially more effective than previously thought" (p. 1686).

In summary, there is considerable research evidence suggesting that leadership training can lead to positive work outcomes and this has been widely accepted in the literature (e.g., Aguinis & Kraiger, 2009; Avolio et al., 2009; Collins & Holton, 2004; Lacerenza et al., 2017). However, we argue that in the absence of a rigorous examination of the research designs employed in this research, and their ability to establish causal inferences, this conclusion may be premature. Next, we describe issues concerning the demonstration of causality in leadership training research and the ability of various research designs to deal with these issues.

3. Determining causality in leadership training studies

3.1 The endogeneity problem

Endogeneity is a technical econometric label that refers to a suite of well-known, but underappreciated and rarely addressed, model misspecifications that can bias parameter estimates (Antonakis, Bendahan, Jacquart, & Lalive, 2010; Hughes, Lee, Tian, Newman, & Legood, 2018). Endogeneity biases fall into three broad categories. First, models often omit variables that provide alternative or additional explanation for the relationship of interest. Second, models often fail to take account of simultaneity, that is, when the dependent variable and one or more of the explanatory variables are jointly determined. In such cases, results from

regression analysis might confound these relationships because the direction of causality most likely runs in both directions (e.g., X→Y→X). Third, estimates can be biased due to measurement errors such as scales that assess only part of a construct (Hughes, 2018) or common method biases (Podsakoff & Podsakoff, 2019).

In technical terms, endogeneity refers to an instance when a measure of a predictor variable (in our case leadership training) is correlated with the error term of the outcome variable (see Antonakis et al., 2010) or as Hughes et al. (2018, p. 558) describe "… an endogenous predictor is related to the measured outcome variable in two or more ways, usually in the way theorized (e.g., as a meaningful cause), but also in some unanticipated way(s) (e.g., common method bias, reciprocal effects, relationship with a common cause)". In leadership training research the 'unanticipated ways' that the predictor could explain the outcome, are many (e.g., confounding due to non-randomization).

Endogeneity biases can be very problematic for estimating training efficacy, because it is difficult to establish the degree to which the estimate represents the theorized relationship (i.e., the causal effect of the training) or the unanticipated relationship (i.e., endogeneity biases). Fortunately, there are a number of methods that can be used to reduce or remove endogeneity biases, such as the use of instrumental variables. Because these methods have been covered extensively elsewhere, we will not repeat them here. Instead, we refer readers to Antonakis, Bendahan, Jacquart and Lalive (2014) for an introduction to these methods, to Antonakis et al. (2010) for a more technical review, and to An, Meier, Ladenburg and Westergård-Nielsen (2019) for an application with leadership training.

Despite prominent and legitimate concerns regarding endogeneity some, however, think that the endogeneity biases can never be fully removed and thus we must learn to 'live with them'. For example, Van Lent (2007) argues that because theory is never likely to be complete enough to produce fully specified models and good measures are rare, there is little that one can do to mitigate endogeneity. Instead, Van Lent argues that researchers should utilise the best

design possible, even if imperfect (e.g., omitting a known predictor due to time constraints). In doing so, researchers should explicitly defend their choice of misspecification, discuss results accordingly, and consider how the results might have changed had the researcher chosen a model with different misspecifications (e.g., include additional predictors but use shorter measures).

Our view is somewhere in between. Endogeneity is a serious concern and all efforts should be made to reduce misspecification caused by endogeneity biases. Randomized experimental designs should be used whenever appropriate. However, many opportunities to study leadership training in the field will not allow for the use of randomized experiments because such designs raise difficulties, such as, participant recruitment, appropriate randomization, and ethical concerns that can be difficult, sometimes impossible, to overcome. When compromise in study design is unavoidable, we do not believe that researchers should immediately abandon empirical testing, because even if endogeneity is present, studies can serve as useful explorations that can inform future research.

3.2 The minimum conditions necessary to infer causality in leadership training research

Below we draw upon the large literature on causal inferencing (e.g., Antonakis et al., 2010; 2014; Morgan & Winship, 2015) to briefly describe the three major conditions needed to infer some degree of causality in relation to leadership training research. In other words, these are the minimum conditions necessary to infer that it is the leadership training that leads to, or causes, changes in outcomes (see Antonakis, et al., 2010 for a discussion of these in relation to leadership research).

First, *x must precede y temporally* (i.e., the leadership training needs to occur before post-test measures). This is likely to be the easiest of the three conditions. By definition, post-test measures are obtained after the leadership training has occurred. However, the necessary time-period between the leadership training and post-test measurement is a much more difficult issue to determine. It is necessary to consider issues around training transfer (i.e., the process by

which new knowledge is manifested on outcomes, Gilpin-Jackson & Bushe, 2007) and temporal design assumptions (i.e., time-periods between the leadership training and outcomes measurements, Fischer, Dietz, & Antonakis, 2017). The time-periods chosen should reflect the theoretical content of the training and correspond to the expected time taken for (a) the training to be manifested in the predicted changes to leaders' behaviors, (b) leaders' behaviors to have a reliable impact upon team members' behaviors and, (c) changes in team members' behaviors to affect work-related outcomes (such as performance). These decisions should be based on the theoretical content of the training and the mechanisms by which the training content will affect the outcomes. If inappropriate time periods for measurement in relation to the training are made, then inappropriate conclusions may occur (Gilpin-Jackson & Bushe, 2007).

Second, *x must be reliably correlated with y* (i.e., the leadership training reliably predicts post-test measures). Examination of the statistical implications of determining causality is beyond the aims of this paper (for reviews see Gangl, 2010; Heckman, 2005; Morgan & Winship, 2015). However, it is worth noting one main issue. In testing the relationship between leadership training and outcomes, researchers generally use statistical techniques that examine associations (e.g., correlation, regression, propensity scores). However, there is a clear difference theoretically and statistically between association and causation (Pearl, 2009).

Associations show that two factors share a joint distribution of observed variables. Association is based on a statistical probability judgment (if you get A then you tend to get B or in this case leadership training is associated with positive outcomes). Association is based on the distribution of data and leads to inferences of probability. Causation must go further and not only determine probability in a static situation but also understand the effect of change as a cause of future events. Association and causation are not good 'bed-fellows' as Pearl (2009) states there "… is nothing in a distribution function to tell us how that distribution would differ if external conditions were to change - say from observational to experimental setup - because the laws of probability theory do not dictate how one property of a distribution ought to change

when another property is modified" (p. 99).

Based on these arguments many scholars note that adequate testing of causation cannot be achieved by probability calculus but needs new mathematical approaches. One such approach could be the use of Bayesian statistics (Kruschke, 2013). Bayesian statistics moves away from estimating the size of an effect (sometimes referred to as the 'frequentist interpretation') to testing for the existence of an effect. One of the many problems with estimating effects is that there is an *a priori* assumption that there is, in fact, an effect to be estimated and that this can be represented in a probabilistic distribution. Testing for an effect should be the precondition before estimating parameters (Jeffreys, 1961). Bayesian statistics provides a mathematical way to combine prior beliefs and evidence to produce posterior beliefs. Bayesian hypothesis testing is conducted by comparing the predictive adequacy of two competing statistical models, namely the prior (probability distribution expected based on prior beliefs or theory) and posterior (probability distribution that represents updated beliefs about the parameter post data collection) distributions (for the benefits of Bayesian statistics see Wagenmakers et al., 2018a,b).

Third, the *relation between x and y must not be explained by other causes* (i.e., post-test measures are due to the leadership training and not to other factors). This is the most difficult step to determine. To answer this, it would be useful to refresh ourselves on the meaning of causality. Heckman (2005, p. 1) nicely describes it as "Holding all factors save one at a constant level, the change in the outcome associated with manipulation of the varied factor is called a causal effect of the manipulated factor". The only way to hold everything else constant except the manipulated variable is to have an identical condition that does not contain the manipulated variable. This is referred to as the control/counterfactual condition and in leadership training would be represented by a non-training condition.

A formal mathematical framework for understanding causal inferences is the Rubin Causal Model (RCM, Imbens & Rubin, 2008; Rubin, 1974, 1975). In essence, they argue, "no causation without manipulation" (Rubin, 1975, p. 238) of the independent variable (i.e.,

obtaining outcomes for conditions with and without the training). By doing this, two critical causal questions can be addressed (a) if those that had received the treatment, in fact, did not receive the treatment - what would be observed on *y* and, (b) if those that had not received the treatment, in fact, do receive the treatment - what would be observed on *y*. The second part of the RCM specifies that these casual questions can only be answered when conditional equivalence is achieved. That is, when the assignment to conditions is random then one can assume that all potential confounding variables are equally distributed between the conditions (i.e., training and control) and cannot account for differences in outcomes. If conditional equivalence is achieved, it logically follows that the post-test outcomes in the control condition would reflect what individuals in the training condition would have achieved had they not been trained (a) and the post-test outcomes in the training condition would reflect what those in the control condition would have archived had they been trained (b).

3.3 Evaluating leadership training research designs

In this section we examine how robust the most frequently employed study designs in leadership training research are to a number endogeneity biases and model misspecifications that can bias parameter estimates and render studies uninformative. In terms of these threats to study validity, we draw upon the classic analysis by Campbell and Stanley (1963). For a brief description of how each of the threats relate to leadership training, see Table 1 (see also Martin, Epitropaki, & O'Broin, 2018; Podsakoff & Podsakoff, 2019). We do not describe these threats in further detail but consider how robust each research design is to them.

<Table 1 about here>

We examine eight specific research designs from three broad categories, namely, intervention only designs, independent groups designs, and independent groups with repeated measures designs. There are many more research designs we could consider (such as, regression discontinuity designs, Shadish & Cook, 2009; and alternatives to randomized control studies, West, 2009) but we focus on these eight because they have been the most frequently employed

in leadership training research. For the purpose of nomenclature, we represent these designs using the following symbols:

- 'O' refers to outcomes that are expected to be affected by the training and can be obtained before (pre-test) and/or after (post-test) the training. Outcome measures can be obtained via various methods (e.g., survey data, observations, performance scores) and sources (e.g., leaders, team members, clients).
- 'X' refers to the leadership training that is predicted to affect O.
- 'C' refers to the counterfactual or control condition, in which participants receive no training (classic control) or receive non-leadership training (active control).
- '–' is used to represent the passage of time, which can vary between and within conditions.

Table 2 provides a summary analysis of the ability of each of the eight commonly employed designs to estimate and control for the threats to study validity described in Table 1. Below, we describe each design in further detail in ascending order starting with the weakest designs.

<Table 2 about here>

Intervention only designs. These designs are characterized by examining participants that only receive the leadership training and do not include a control condition (i.e., those that do receive training). Three designs can be identified in this category. The first, *one shot case study design*, is the most basic and only includes the assessment of outcomes once after the leadership training (*X – O*; e.g., Gilpin-Jackson & Bushe, 2007). One shot case studies are most open to validity threats because the lack of pre-training scores renders the design highly vulnerable to *history* (e.g., outcomes may have been increasing over time independent of the training), *maturation* (e.g., leaders and/or team members' competencies improve over time with increased experience), *selection* (e.g., 'better' managers are chosen to receive the training), and *mortality* (e.g., poorer leaders or team members leaving the company during the study period) effects.

Thus, one shot case study designs yield very weak evidence that changes in the outcomes were caused by the training and so should not be used to inform policy recommendations.

A development of the one shot case study would be to include pre-test measures; this is referred to as a *repeated measures design (O – X – O*; e.g. Arthur & Hardy, 2014; Biggs, Brough, & Barbour, 2014). Repeated measures designs can deal with some of the threats in the one shot case study design, such as *mortality* effects (e.g., managers leaving the organization or changing jobs and therefore exit the study). Mortality effects can be estimated by examining the pre-test scores for those that remain versus those that leave the company during or after the training. If there are no differences, one could assume that mortality is not an issue. Also, pre-testing allows gain scores to be calculated (i.e., post-test minus pre-test) that can quantify the impact of the intervention. However, although repeated measures designs address some threats, they introduce others, such as *testing*, which is a particular problem in leadership training because the pre-test could unintentionally alert participants to aspects of the leader's behavior, that they had not previously considered, and this affects their subsequent post-test scores (independently of the training).

The repeated measures design typically has one set of pre-test and post-test measures. However, it is possible to have multiple pre-test and post-test assessments of outcomes and this is referred to as a *time-series design* (*O – O – X – O – O*; e.g., DeRue, Nhargang, Hollenbeck, & Workman, 2012). As discussed earlier, choosing time periods between assessments that are optimally designed to capture the target effects is crucial, and it is likely that outcomes will be influenced at different rates (see discussion above on establishing the first criteria for causality).

There are many variants of time-series designs (in terms of the frequency of pre-test and post-test measures) but the aim is to show that the outcomes change directly as a result of the training in a manner that is different to other time periods. The 'ideal' pattern would be to show (a) a consistent level in outcomes on all pre-test time periods, (b) an improvement (this could be an increase or decrease depending on the outcome) immediately following or soon after the

training, that (c) remains consistent or increases/decreases as predicted at later post-test time periods. Although this is the ideal pattern, it is rare in practice. Nevertheless, many other patterns can also indicate benefits of leadership training. For example, outcomes could be improving at pre-test periods but show greater increases after the leadership training that is maintained on post-test periods.

Overall, time series designs can deal with many threats to validity except in relation to history. It could be that changes in outcomes reflect some underlying changes in the organization or context that is independent of the leadership training. Multiple measurement periods pre- and post-intervention, combined with predicted outcome changes following the intervention, could mitigate, but not completely remove, this concern.

Independent groups designs. The earlier discussion on establishing causality shows the need for research designs to include a control condition to establish what would happen if participants had not received the leadership training. A number of research designs do this by having at least two conditions: leadership training and non-leadership training (control). We discuss the nature of the control condition in a later section. In order to reap the benefits of independent groups designs, participants must be allocated to the two conditions in either a random or matched (i.e., non-random but attempts to make groups equivalent in terms of potentially key confounding variables). If the allocation is random, then the research design is often referred to as a *laboratory experiment* (if the investigators also have total control over the participants' environment) or as a *field experiment* or *randomized control study* (if there is not control over the environment, as is typical in organizational studies).

The most basic *independent groups* design assesses outcomes only during the post-test (*X – O, C – O*). When participants are randomly allocated (e.g., Moses & Richie, 1976), then the basic independent groups design is very powerful, and can attenuate all eight major threats to validity (see Table 1). However, when participant allocation is non-random (e.g., Ladegard & Gjerde, 2014), the design is susceptible to a number of threats, such as, selection, mortality, and

selection by maturation interaction. One disadvantage of the basic independent groups design, regardless of the participant allocation procedure, is that the lack of pre-test outcome assessments prevents the calculation of the nature or absolute magnitude of changes due to the training. For example, it is possible that both groups worsen their outcomes post-training but that the reduction is lower in the training condition. The inclusion of pre-test measures allows one to address this issue and determine absolute magnitude of change, which is important when calculating the return on investment of the training.

Independent groups with repeated measures designs. When independent groups also have pre-test assessments, they are referred to as *independent groups with repeated measures* designs (*O – X – O*, *O – C – O*). When participants are randomly allocated (e.g., Chernis, Grimm, & Liautaud, 2010; Yeow & Martin, 2013), the independent groups with repeated measures design is well-placed to estimate and attenuate all eight threats to internal validity, and thus, is often considered to be the best design to examine leadership training. There is an additional advantage of the use of pre-test measures over previous designs. Condition equivalence which is assumed in the independent groups design, when there is condition randomization and sufficient sample size, can be tested and confirmed if pre-test scores are similar across conditions.

However, there remain two notable concerns depending on the execution of the independent groups with repeated measures design. First, non-random participant allocation (e.g., Elo, Ervasti, Kuosma, & Mattila-Holappa, 2014) renders these designs vulnerable to some of the threats to validity (e.g., selection and selection by maturation interaction) especially if participants in the training and control conditions are non-equivalent in terms of important confounding variables (e.g., tenure, performance). Matching techniques, such as propensity score analysis, where equivalent participants can be identified in each condition on key criteria can re-establish conditional equivalence. Second, testing effects might occur when the pre-test measures assessed for those in the training condition differ from those in the control condition.

In such cases, the pre-testing might differentially alert respondents to key causal variables that distorts post-test scores. The simple recommendation here is to ensure identical pre-test and post-test assessments for training and control groups.

The last design, the *Solomon Four Group Design* (Solomon, 1949), extends the independent groups with repeated measures design by including two additional conditions (training + post-test and control + post-test; $O - X - O$, $O - C - O$, $- X - O$, $- C - O$) (e.g., Birkenbach, Kamfer, & ter Morshuizen, 1985; Holdnak, Clemons, & Bushardt, 1990). As with the previous design, it is important that leaders are randomly allocated into each of the four conditions and that measurement periods coincide. Like the independent groups with repeated measures design, the Solomon design can deal with all the threats to internal validity. However, the Solomon design has an additional advantage in that it can deal with a threat to external validity (i.e., generalizability of findings), namely, the testing by condition interaction. Testing by condition interactions occur when the pre-test measures affect sensitivity to the leadership context which affects post-test measures (testing) and this occurs in both the training and non-training groups (conditions) although to a different degree. If this occurs, then the results for a pre-tested group would not be representative to a non-pre-tested group from which the participants were drawn making it difficult to generalise the results (which is important from a practical perspective). If differences in post-test outcomes between the leadership training and control conditions are equivalent, irrespective of completing the pre-test, then one can conclude that the effect is robust to the testing by condition bias. Although the Solomon Four Group Design is the most robust research design to examine leadership training, very few studies have employed it (Birkenbach et al., 1985).

4. Conceptual approach to the review

This section builds upon the previous section concerning the conditions necessary to establish causality and identifies six critical considerations in leadership training that we use to evaluate the literature. Below, we describe each of the six critical considerations.

4.1 Control condition

Control conditions are necessary in order to establish casual identification (Rubin, 1975). Most commonly, participants in the control condition receive no training. Non-training control conditions provide an unfair comparison because the training group can exhibit significant effects due to placebo, expectancy (i.e., participant unconsciously affects the expected outcome), or Pygmalion (i.e., training procedures increase self-efficacy, and therefore outcomes, independently of the training) effects (Cooper & Richardson, 1986; Eden, 1992). In addition, a leadership training intervention has two potentially active components that can affect outcomes: the *training content* (the focus of investigation) and the *training process*. The training content refers to the knowledge and skills that are developed within the leader via the developmental activities. The training process refers to everything involved in the intervention other than the content (e.g., selection to a course, interaction with other colleagues, interaction with trainers, the physical location, break from the workplace). Participants in the non-training condition receive neither the training content nor the training process and therefore it is impossible to identify which causes change in post-test outcome assessments. Thus, researchers should ideally compare leadership training with a non-leadership training condition (or alternative leadership-training condition) that would mean the control condition is matched in terms of placebo, expectancy, training process effects (e.g., Chambless & Hollon, 1998). Thus, we recommend the use of randomized experimental designs that include a non-leadership (or alternative leadership) training control condition (e.g., Boies, Fiset, & Gill., 2015).

The nature of the control condition might also influence participants' motivation to complete pre-test and post-test assessments. For managers, and their team members, in the leadership training condition, providing outcome data would seem a natural part of the development process. Indeed, such data might be used in the training to inform participants of their leadership behaviors (e.g., 360 degree feedback). The motivation to participate in data collection would likely be reduced for those in the control condition, because the utility of the

data collection would be unclear, increasing the risk of lower response rates or incomplete data. To overcome this problem, all managers could be trained in a time-lagged manner, meaning that the control condition consists of managers who have not been trained but know that they will be trained in the future (i.e., a 'waitlist' control, e.g., Neck & Manz, 1996). Participants in a waitlist control condition would likely have similar motivation to those in the leadership training condition to engage in data collection.

4.2 Sample representation

Sample representation (i.e., the sample is representative of the population from which it is drawn) is an important, but relatively ignored concern. A pragmatic problem in obtaining sample representation is that the researcher typically has no or little control regarding who is selected for training. For organizations, training costs are considerable (both financially and time-wise), which sometimes affects the selection process. Sometimes, managers who are available, whose absence is less disruptive to performance, or those considered 'rising talent' are selected. Such selection strategies increase the likelihood of group differences that can act as confounding variables (the *selection bias*). This affects both hypothesis testing (e.g., by leading to restriction of range) and renders the results of the study non-generalizable to the broader population (i.e., other managers in the organization). Collectively, the effects of selection biases can produce misleading results and inappropriate policy recommendations. To avoid selection biases, studies should include all relevant managers (e.g., managers within a set grade or position), use a matching procedure, or draw a random sample of relevant managers. If randomization is conducted correctly, with a sufficient sample size, then potential confounding variables are likely to be randomly distributed between the population and study sample, meaning the results can be generalized to the population.

Claims of sample representativeness depend on whether those selected actually do the training. If participation is voluntary there may be many contextual factors (e.g., work demands, managerial cover for absence, illness, personal preference) that might affect training attendance.

All these factors could lead to biased and unrepresentative samples. Papers tend to be vague when reporting sample representativeness and in particular reasons for training non-attendance. One potential way to avoid this is to make training attendance involuntary. The issue of training voluntariness was explored in the Lacerenza et al. (2017) meta-analysis across Kirkpatirck's (1959) evaluations of training. They found mixed effects. Whereas voluntary participation had a bigger effect than involuntary training on transfer (i.e., using trained skills and abilities in the workplace), it had a negative effect for 'results' (i.e., improved organizational performance objectives), and no effect on 'learning' (i.e., changes in knowledge and skills). Although the 'jury is out' on the impact of training voluntariness on training outcomes, it is an unresearched factor that could severely bias the training population. One might argue that condition randomization should negate such a problem but it does not. The problems of voluntariness affect the training but not the non-training condition and could lead to conditional non-equivalence.

4.3 Condition randomization

The most robust research designs employ random allocation of leaders into training and control conditions because randomization greatly increases the likelihood that the leaders in each condition are equivalent or at least that there are no systematic leader differences that might explain differences in outcomes. Randomization is effective when sufficiently large sample sizes are obtained. Unfortunately, the adequacy of sample sizes tends to be assumed rather than explicitly examined, despite the availability of methods to calculate appropriate sample sizes (e.g., Rutterford, Copas, & Eldridge, 2015).

The problem of achieving randomization are not unique to leadership training interventions, with general agreement that it is difficult to randomly assign participants in the field (Evans, 1976), meaning that "… the methodological requirements of traditional experiments fail to mesh with the realities of life in organizations" (Lawler, 1977, p. 577). As an alternative to randomization, matching procedures based on criteria that are linked to training

success can, under specific situations, create conditional equivalence (Stuart, 2010).

4.4 Condition independence

Even if condition randomization does occur, there are a number of additional problems that might arise. One such problem concerns the independence of managers and team members between conditions and the potential of *carry-over effects* (in some research areas this is referred to as 'interference'). If managers in the training and control conditions are from the same organizational site, it is likely that trained managers (and their team members) interact with, and therefore potentially affect, control managers (and their team members). To avoid cross-condition contamination, researchers can select managers for each condition from different regional sites (e.g., Paul, Robertson, & Herzberg, 1966) but this can increase the likelihood of regional selection biases. Selecting managers from different sites in this way provides *between* condition independence (i.e., leaders/followers from one condition do not interact with those from another condition) but not *within* condition independence (i.e., leaders/followers within a condition do not interact with others from the same condition). Since most training involves participants being trained together in groups, it would be difficult to ensure both within and between condition independence. Due to this, we focus on between condition independence, which ensures that there is no cross-condition contamination or interference between those in the training and non-training conditions.

4.5 Temporal design

The timing of pre-test and post-test measures is critical in order to capture leadership training related changes in outcomes. To do this, researchers need to understand the training transfer process (Grossman & Salas, 2011). Russon and Reinhelt (2004) note a lack of consideration for the timing of outcome measurement, with post-training measurers often obtained soon after training completion, perhaps because of the pressure from funders to show positive effects as soon as practical. Another factor that could encourage this situation is the concern researchers might have to collect data before changes to organizational context (such as

new working practices, changes to management) that might interfere with potential benefits from the training or inhibit future data collection.

Decisions concerning the timing of measurements could be guided by Day and Sin's (2011) distinction of the visibility of the leadership development process at the surface, meso, and foundation levels. Leadership training that aims to influence behaviors at these different levels will need different time-periods in order to capture the work outcomes. At the surface-level (visible), the intervention might target specific leader behaviors that are easily adapted, meaning post-training assessments could be taken relatively soon after completion (such as charismatic leader training, e.g., Antonakis, Fenley, & Liehti, 2011; transformational leadership training e.g., Barling et al., 1996; LMX training, e.g., Graen, Novak, & Sommerkamp, 1982). At the meso-level (less visible), the predicted change might target identity and self-regulation processes and therefore more time is needed for the training to impact the outcomes, meaning post-training assessment should be taken some time later (e.g., leader identity, Miscenko, Guenter, & Day, 2017). Finally, at the foundation-level (invisible), the predicted changes might target major reshaping of the leader's self-concept and/or behavioral repertoire, both of which require considerable time to become detectable, meaning post-training assessment should be taken long after the training (e.g., adult development processes, Day & Sin, 2011). Studies that use a time-series design (multiple measurement time periods) should time the measurements according to the leadership development level.

Similar considerations pertain to pre-test measures, which should be assessed in a time-appropriate manner. For example, if job performance is an outcome, timing of the pre-test (and post-test) should reflect the unit of performance relative to the performance cycle for that job. So, a sales person on a monthly sales target could have pre-test measures much nearer in time to the training compared to a marketing manager whose projects might take several months to complete.

4.6 Author involvement

Who is responsible for the training design and delivery is an important issue because it can affect leadership training (see Lacerenza et al., 2017). Training can be conducted in-house (most typically via HR specialists) or via external training agencies, but the key question pertains to the involvement of the research team and the relationship that they have with the client organization. If the research team design and/or deliver the training, they are likely to have a vested interest in the outcome, because successful training can yield many benefits (e.g., payment, publications, repeat business, esteem). As a consequence, researchers' involvement in training design and delivery represents a potential conflict of interest (Chivers, 2019) that can be considered a demand characteristic (i.e., a factor, other than the research design, that might unwittingly affect the outcomes; Orne, 1962). For example, without realising the research team may deliver the leadership training with more gusto, compared to those that are not involved in training design, which improves outcomes via processes other than the training content. A desire to provide support for the training programme can also manifest in questionable research practices in processes such as data collection, analysis, reporting findings, and conclusions drawn (John, Loewenstein, & Prelec, 2012). For example, researchers can write papers with a narrative 'story' focussed on select outcomes that 'work', perhaps omitting, altogether, outcomes that fail to 'work'. Equally, researchers can add or remove covariates, include or exclude outliers, and so on, depending upon on how they shape the story. All of these decisions, often made without malicious intent, serve to distort the research record and produce a misleading and potentially harmful literature. Thus, if researchers are involved in any aspect of, and/or profit from, the training intervention this should be stated explicitly as a potential conflict of interest and acknowledged as a potential threat to the validity of the findings.

5. Review of leadership training literature

In this section, we provide a review of the leadership training literature in terms of the six critical considerations for research design identified in the preceding section.

5.1 Method: Paper selection and coding criteria

Our sample is drawn from the published studies analysed by Lacerenza et al. (2017) in the most recent meta-analysis of leadership training. Lacerenza at al. (2017) primary study research designs into three categories: single group repeated measures (208 studies, i.e., pre-test and post-test outcomes for leadership training, with no control condition), independent groups (62 studies, i.e., intervention and control condition with post-test outcomes) and, independent groups with repeated measures (58 studies, i.e., intervention and control conditions with pre-test and post-test outcomes). Although the very wide confidence intervals surrounding the point estimate (i.e., best estimate) of training efficacy across these study designs overlapped, there was a notable difference in the point estimates. Specifically, the average effect (Cohen's d corrected for criterion unreliability) for independent groups was $\delta = .78$, for single group repeated measures was $\delta = .76$, and for independent groups with repeated measures was $\delta = .48$. Thus, as study design rigor increased, the average effect decreased.

Since we are specifically interested in research designs that employ a control condition because control conditions are necessary to determine causal identification (Rubin, 1974), we focused on two sets of published papers: independent groups ($n = 30$) and independent groups with repeated measures ($n = 32$). The Lacerenza et al. (2017) meta-analysis used a strict set of criteria for inclusion to meet their meta-analytic needs, this means that some studies in the literature were excluded (see the Lacerenza et al., 2017 paper and the supplementary information for selection criteria and the studies included). However, since there is a large overlap in samples between the Lacerenza et al. (2017) and Avolio et al (2009) reviews, one can assume that the former is a fair representation of the existing literature.

We coded each paper to reflect the six core issues described above related to causality:

1. *Control condition* (no training, waitlist, non-leadership training).
2. *Sample representation* occurs where either all leaders or a random sample of leaders is drawn from the available population (yes, no).

3. *Condition randomization* between the training and control groups (yes, matched, no). We also recorded the sample size of each condition and also if there were reliable difference on pre-test scores (yes, no).
4. *Condition independence* between the training and control conditions in that leaders and their team members could interact with those in the control condition (yes, no).
5. *Temporal design* in terms of a clear justification of timing of pre-test and post-test measures to capture training transfer (yes, no). We recorded the time periods between pre-test, leadership training and post-test (in weeks). We also coded whether pre-test and post-test measures were taken at the same calendar time (yes, no).
6. *Author involvement* in either content development or training delivery (yes, no).

If it was not possible to make a judgement, or the information was not available, then this was recorded as a 'no coding'.

5.2 Results and discussion

Table 3 contains a summary of the codings in relation to the six core issues identified above.

<Table 3 about here>

Control condition. The majority of studies employed a non-training control condition (*n* = 54, 87%), meaning they were unable to partition the effects of training content and process. To do this, a non-leadership (or alternative leadership) training control condition in which leaders are exposed to the training process is required. Six studies used a non-leadership training control condition (e.g., Smith, Montello, & White, 1992). Finally, eight studies (13%) used a waitlist control condition, in which participants knew that they would receive the leadership training in the future (e.g., Neck & Manz, 1996). Waitlist conditions most commonly occur when managers are trained in batches (or cohorts) due to resource constraints. Waitlist conditions are useful because they control for possible effects of variations in participant motivation to participate in data collection.

Sample representation. In over 70% of studies (n = 44) it was not possible to determine if the study sample was representative of the population it was drawn from because little or no information regarding sample selection was presented. In studies where sample representation could be established (n = 18, 29%) this was achieved by using the entire population of managers (e.g., Birkenbach et al., 1985) or by drawing a random sample from it (e.g., Tharenou & Lydon, 1990). However, in studies using a random sample from the population, the criteria for determining randomization was not clear (e.g., the selection percentage).

Condition randomization. Five studies used a matching procedure (i.e., allocated participants to conditions based on pre-existing criteria to determine equitable groups). However, in four of these studies the matching procedure was based on factors that did not appear to be related to training transfer such as biographical data (e.g., Carron, 1964) or similar work sites (e.g., Moffie, Calhoon, & O'Brien, 1964). The study by McGehee and Gardner (1955) used a matching procedure where participants were allocated to conditions based on their pre-test scores and demographic factors. Assuming the pre-test measures tap into factors that might be related to training success, then this would be considered to be a good technique. However, since there was only one study that did this, we combined the matched studies with those that did not report using randomization.

Overall, twenty-eight studies randomly allocated leaders and thirty-four did not. Condition randomization is the most important aspect to determine causality in research because group members should not vary systematically on any potentially confounding variable. Thus, it is surprising that less than 50% of studies employed random allocation. In total, twenty-three studies employed pre-test measures that allowed for the estimation of the magnitude of pre-test differences between conditions. When there was no condition randomization, nine studies found pre-test differences and seven (43%) did not. However, when there was condition randomization, only one study found pre-test differences and six (86%) did not. This shows the

importance of using condition randomization to establish the crucial criteria of equivalent groups.

To be effective, randomization needs sufficiently large samples to be able to assume that potential confounders are evenly distributed between conditions (Rutterford et al., 2015). Overall, the number of leaders in training conditions (M = 41.11) tended to be slightly larger than control conditions (M = 35.90). However, there was wide variation between studies and it is useful to note (a) more studies had a greater number of participants in the training than control condition (n = 28, 45%) than vice versa (n = 13, 21%) and (b) some studies had very low number of participants with 20% of studies (n = 12) having 10 or fewer participants in a condition.

Condition independence. In over 70% of the studies condition independence between the training and control conditions had not been achieved (n = 44, 71%). This is a problem because one is unable to rule out cross-contamination between the leaders (and their team members) in training and control conditions, which can pose a serious threat to the validity of the study. In cases when condition independence was achieved, it was often due to leaders in different conditions being located in physically separated sites (e.g., Porras, Hargis, Patterson, Maxfield, Roberts, & Bies, 1982). Whilst this might control for cross-contamination, it introduces other possible confounding variables, such as selection bias and site differences.

Temporal design. There was considerable variation in time periods between pre-test and post-test in relation to the leadership training. The situation is made more complicated because the length of training varied in time from one day to several months. However, the mean time (in weeks) from pre-test to training was 3.72 (n = 29) and between training and post-test it was 24.46 (n = 46; the difference in training to post-test between studies with and without a pre-test was similar: 26.13 vs. 22.91). It is also evident that, for the studies that could be coded, in most cases pre-test and post-test measures were not collected at the same calendar time points (n = 28, 67%). In terms of studies justifying the pre-test and post-test time periods in terms of

transfer of training criteria was very clear cut: only one study did this (Steensma & Groeneveld, 2010). Because the vast majority of papers provide no justification (e.g., theory, client/researcher needs, participant availability) for the time periods between the leadership training and collection of outcome data, it is unclear how they were determined and whether they were optimally designed to capture changes in the outcomes.

Author involvement. In only about a quarter of studies it was clear that the authors had no involvement in training design or delivery (n = 15, 24%). Over half the studies reported that the authors were involved with some aspect of the training design or delivery (n = 34, 54%). The potential for demand characteristics affecting the results in these studies, due to conflicts of interest, could be heightened. However, on a positive note, when authors were involved in the study, condition randomization procedures were likely (ns = 24 vs. 10) than when they were not involved in the study (ns = 1 vs. 14). One might suspect that when organizations design leadership training initiatives, they do not consider issues concerning causality and therefore condition randomization might be implemented on the advice of a research team.

6. Additional observations, summary and recommendations

In the final section we consider some additional observations from our review (in relation to training content and time series), summarize some of the issues we have considered in relation to establishing causality in the leadership training literature, and make recommendations to guide future research in the area.

6.1 Training content

Although leadership training content was not an explicit focus of our review, we did notice that most papers included insufficient details to assess what the leadership training involved. Thus, most studies could not be accurately replicated. Further, the training content varied so much that we would be hard pressed to find any two studies that were equivalent in terms of leadership content! For example, studies that aimed to improve leaders' transformational leadership (e.g., Barling et al., 1996; Bellé, 2013; Dvir, Eden, Avolio, &

Shamir, 2012; Parry & Sinha, 2005) varied in length (e.g., half a day, 3 days, 3 months), format of delivery (e.g., workshops, individual sessions), development activities (e.g., role-playing, coaching, talks), and inclusion of additional theoretical content (e.g., goal setting, motivation theory). Without clear replications of effects and equivalence between training programmes, it is difficult to say what is driving any causal effect. This problem is compounded in reviews and meta-analyses that group studies that purport to develop the same leadership concepts, and as a result, potentially misidentify the true causal mechanisms.

Another issue is that in the majority of studies the content of the leadership training was not based on leadership theory but targeted a narrow range of job-specific behaviors. In these cases, it is difficult to determine if the training was directed at leader or leadership development (Day & Dragoni, 2015). To make matters worse, many studies that claim to involve leadership training seemed to have little to do with leadership processes, instead focusing on general managerial skills (see Collins & Holton, 2004). Overall, there needs to be a much clearer understanding in these studies of what is in the training (content and process) and how it relates to theory (see Avolio et al., 2009).

6.2 Time trends

The number of studies per decade (from 1950s to 2014) has been similar (average 8.9/decade, range 3 to 14). It is pleasing to note that most papers have been published in the most recent decade (n = 14, 2010-2014) which, at the time of writing, has not yet been completed. However, one concerning feature was a clear downward trend in relation to condition randomization. Condition randomization was achieved in 55% of earlier studies (1950s-1980s; n = 16 out of 29) and only 36% in those conducted later (1990s-2014; n = 12 out of 33). The fall in the number of studies reporting condition randomization is most notable in the present decade (which is not complete), during which, only 3 (21%) of 14 studies used condition randomization. We should add a note of concern because many recent studies fail to describe the condition allocation procedure (leaving the reader to assume randomization was not

achieved) and do not discuss this as a potential threat in the paper. This, we believe, is worrisome and something we urge future researchers to attend to.

6.3 Establishing causality in leadership training research

Our review shows that research designs differ in their ability to deal with endogeneity biases and model misspecifications that can bias parameter estimates and yield inaccurate results. Overall, randomly assigned control studies meet many of the criteria for determining causality. The Solomon design is perhaps the most powerful although least employed. This is perhaps because it involves four conditions meaning it is labor intensive and requires many participants. However, given the power of this design, we would encourage more use of it when possible.

Research designs, by themselves, do not cure all potential confounding problems. Issues concerning endogeneity, for example, can be reduced but not necessarily eliminated by research design features. In designing studies, researchers need to explore the potential for endogeneity bias in their study and how it might influence the results (see An et al., 2019). If endogeneity biases are sufficiently large this renders the study uninformative and potentially harmful (i.e., by drawing erroneous and incorrect conclusions) and it is best to abandon attempts to statistically quantify the training effects. Although researchers may not be able to deal with all biases in every study they should be explicitly acknowledged when making any claim of causal effects.

Our analysis identified a number of important but underappreciated issues that can undermine claims of causality in leadership training research: control condition, sample randomization, condition randomization, condition independence, temporal design, and author involvement. Our review highlighted that the leadership training literature is vulnerable to all these factors and in some cases to a very high level. While we were unable to find examples of good practice over all six factors, we observed good practice in each of them. For example, some studies employed both non-leadership and non-training control conditions (e.g., Fitzgerald & Schutte, 2010); achieved sample representation (e.g., Ivancevitch, 1982); employed condition

randomization (e.g., Dipietro. 2006); used designs where cross-contamination between conditions was not possible or very unlikely (e.g., Porras et al., 1982); where temporal design was justified (e.g., Steensma & Groeneveld, 2010), and where authors were not involved in the training delivery (e.g., Bozer, Sarros, & Santora, 2013).

Despite some good practice, the general prevalence of sub-optimal study designs that are interpreted as providing clear evidence for causality is high. For example, in over 70% of studies it was not clear if the sample was representative of the population from which it was drawn leading to problems in generalization of findings. Given the crucial role of condition randomization it is surprising that less than half of studies achieve this. Over 50% of studies did not have random procedures making their post-test results vulnerable to a wide range of biases (such as selection, history) especially when they did not have a pre-test. In the absence of condition randomization few studies used matching techniques (see Stuart, 2009, or techniques such as propensity score analysis) to establish greater levels of condition equivalence (for a good example see Neck & Manz, 1996). The importance of condition randomization was shown by the fact that when it occurred there was, as expected, less likely to be differences in pre-test scores (showing that condition equivalence had been achieved). Condition independence is also necessary in order to rule out cross-contamination effects (e.g., managers in training and non-training condition mix and influence each other). However, condition independence could not be established in 70% of studies. It was further surprising that only one study explicitly considered the temporal design. Training transfer issues (Gilpin-Jackson & Bushe, 2007) and temporal design assumptions (Fischer et al., 2017) guided by the theoretical content of the training need to be considered for appropriate conclusions to be drawn (Gilpin-Jackson & Bushe, 2007). Put simply, many studies might be trying to capture changes in outcomes before they can be manifested within the target population. Given the dynamic nature of leadership phenomena, issues related to the nature and the direction of the expected change (e.g., discontinuous, uni-directional or non-linear change) need to be acknowledged (McClean, Barnes, Courtright, &

Johnson, 2019). Finally, in over half the studies the authors had a role in training design or delivery and this was most notable when there was condition randomization. We suspect that might be because of the active involvement of the research team in the leadership training process. Author involvement, which can lead to a conflict of interests and serve to increase potential demand characteristic, needs closer examination (Chivers, 2019).

6.4 Recommendations

We repeat our view that applied studies examining leadership training are difficult to conduct and there is a tension between what researchers wish to do and what they can do. However, we hope these guidelines will serve to stimulate further consideration, inform choices about research design issues, and encourage better reporting of relevant information to aid fair evaluation. These recommendations are focussed on field and quasi-experimental field studies, but are also relevant for laboratory experiments (see Lonati, et al., 2018, for guidance on experimental issues). We group these recommendations around the six crucial conditions for establishing causality employed in the literature review.

Control condition

- The use of control conditions is essential. Studies should employ them.
- No training control conditions are useful, but there is a need to understand the effects of both training content and training process on outcomes.
- Where possible, active control conditions should be employed such as *alternative leadership content* (i.e., more than one intervention that covers different aspects of leadership content, e.g., Dipietro, 2006), *non-leadership content* (i.e., training in areas not expected to affect leadership processes, e.g., Fitzgerald & Schutte, 2010), or *waitlist* (i.e., participants are due to receive intervention in the future, e.g., Neck & Manz, 1996) conditions. Each of these conditions can address the effects of training content and process, and different threats to validity helping to establish causal identification.

- Test all potential threats to validity/biases inherent in the research design before testing study hypotheses.

Sample representation

- Sample representation is important to achieve unbiased study estimates and to ensure generalizability to the population.
- Studies need to report how participants are selected for the study and acknowledge any potential selection biases.
- When population size is large, random selection of participants should be followed.
- Sample representation is important for generalizability of findings and informing policy decisions.

Condition randomization

- Condition randomization is crucial for establishing conditional equivalence.
- Randomization procedures require sufficient sample sizes to assume equality of potential confounders between conditions. There are mechanisms to determine appropriate sample sizes for specific research designs.
- Matching procedures can create conditional equivalence. Matching criteria should reflect theoretically identified factors that are likely to impact on the transfer of training. Adequate tests should be conducted to ensure equality at pre-test.
- If randomization is not possible, then pre-test differences between the conditions must be examined on factors that theoretically would affect the leadership process and post-test outcomes. Propensity score analysis can be useful in this context.

Condition independence

- Where possible, there should not be the potential for cross-contamination effects (i.e., leaders and team members interact with those in other conditions).
- If condition independence is not possible, then it should be acknowledged and considered a potential confounding variable.

- The amount of condition independence should be assessed (e.g., more likely for some managers than others) and built into model testing procedures.

Temporal design

- Describe the transfer of training process i.e., the process by which, and when, training content is expected to lead to improvement in the outcomes.
- Assess appropriate outcomes to capture the transfer of training process (both mediators and outcomes).
- Based on the transfer of training process, determine the most appropriate time periods for pre-test and post-test measures.
- Design different measurement time periods to capture optimally different types of outcomes.

Author involvement

- Acknowledge the role of the research team in the leadership training (e.g., training design and delivery). If the authors are involved in any aspect and/or profit from the training intervention, this must be stated explicitly as a potential conflict of interest.
- Describe the relationship between the research team and client organization and acknowledge any potential conflicts of interest.
- If the research team are involved in the leadership training, then they should ensure that outcome data is collected and analysed by independent individuals (i.e., those who do not have a conflict of interest).

7. Concluding postscript

The above recommendations might lead the reader to conclude that we believe there is an 'ideal' leadership training study that addresses all our recommendations. If there is, we have failed to find one and given the complexity and 'messiness' of 'real world' settings we are not optimistic that one will ever be conducted. However, we agree with Shaver (2019) that the nature of leadership studies and the data they produce mean that we can only build a thorough

understanding through an accumulation of evidence generated through studies with various strengths and weakness. Thus, all studies have the potential to contribute to the literature, regardless of their ability to demonstrate causality, but in different and proportionate ways. Indeed, combining studies across different research designs and/or ignoring the types of issues noted above, as is typical of meta-analytic reviews, does not give a true estimate of the effectiveness of leadership training. Being able to evaluate studies' robustness to endogeneity effects and ability to determine causality should allow reviewers to accurately weight their contribution. We would argue that a large effect size in a poorly designed study (which could be accounted by many confounding factors) is less meaningful, theoretically and practically, than a smaller effect size in a well-designed study. It would be helpful if future meta-analyses could focus on the research design strengths of each study as moderators of the impact between leadership training and outcomes. Of course, we wish that this paper stimulates better designed leadership training studies that are more able to determine causality but we hope it also stimulates a better understanding of leadership training research and a greater appreciation of the different contributions each study design can provide. Given the potential for leadership training to significantly improve a wide range of work-related outcomes (such as psychological well-being and performance), or to waste a large amount of time and money, it is an important endeavor for scholars to examine the utility of leadership training.

References

Aguinis, H., & Kraiger, K. (2009). Benefits of training and development for individuals and teams, organizations, and society. *Annual Review of Psychology*, *60*, 451-474.

An, S. H., Meier, K. J., Ladenburg, J., & Westergård-Nielsen, N. (2019). Leadership and job satisfaction: Addressing endogeneity with panel data from a field experiment. *Review of Public Personnel Administration*, 1-24.

Antonakis, J., Bendahan, S., Jacquart, P., & Lalive, R. (2010). On making causal claims: A review and recommendations. *The Leadership Quarterly*, *21*, 1082–1120.

Antonakis, J., Bendahan, S., Jacquart, P., & Lalive, R. (2014). Causality and endogeneity: Problems and solutions. In D. V. Day (Ed.). *The Oxford handbook of leadership and organizations* (pp. 93–117). New York: Oxford University Press.

Antonakis, J., Fenley, M., & Liechti, S. (2011). Can charisma be taught? Tests of two interventions. *Academy of Management Learning & Education*, *10*, 374-396.

Arthur, A. C., & Hardy, L. (2014). Transformational leadership: A quasi-experimental study. *Leadership & Organization Development Journal*, *35*, 38-53.

Avolio, B. J., Avey, J.B., & Quisenberry, D. (2010). Estimating return on leadership development investment. *The Leadership Quarterly*, *21*, 633-644.

Avolio, B. J., Reichard, R. J., Hannah, S. T., Walumbwa, F. O., & Chan, A. (2009). A meta-analytic review of leadership impact research: Experimental and quasi-experimental studies. *The Leadership Quarterly*, *20*, 764-784.

Barling, J., Weber, T., & Kelloway, E. K. (1996). Effects of transformational leadership training on attitudinal and financial outcomes: A field experiment. *Journal of Applied Psychology*, *81*, 827-832.

Bellé, N. (2013). Leading to make a difference: A field experiment on the performance effects of transformational leadership, perceived social impact, and public service motivation. *Journal of Public Administration Research and Theory*, *24*, 109-136.

Biggs, A., Brough, P., & Barbour, J. (2014). Relationships of individual and organizational support with engagement: Examining various types of causality in a three-wave study. *Work & Stress*, *28*, 236-254.

Birkenbach, X. C., Kamfer, L., & Ter Morshuizen, J. D. (1985). The development and the evaluation of a behaviour-modelling training programme for supervisors. *South African Journal of Psychology*, *15*, 11-19.

Boies, K., Fiset, J., & Gill, H. (2015). Communication and trust are key: Unlocking the relationship between leadership and team performance and creativity. *The Leadership Quarterly*, *26*, 1080-1094.

Bozer, G., Sarros, J.C., & Santora, J.C. (2013). The role of coachee characteristics in executive coaching for effective sustainability. *Journal of Management Development*, *32*, 277-294.

Burke, M. J., & Day, R. R. (1986). A cumulative study of the effectiveness of managerial training. *Journal of applied Psychology*, *71*, 232.

Campbell, D. T., & Stanley, J. C. (1963). Experimental and quasi-experimental designs for research. *Handbook of research on teaching*, 171-246.

Carron, T. J. (1964). Human relations training and attitude change: A vector analysis. *Personnel Psychology*, *17*, 403-422.

Chambless, D. L., & Hollon, S. D. (1998). Defining empirically supported therapies. *Journal of Consulting and Clinical Psychology*, *66*, 7-18.

Cherniss, C., Grimm, L. G., & Liautaud, J. P. (2010). Process-designed training: A new approach for helping leaders develop emotional and social competence. *Journal of Management Development*, *29*, 413-431.

Chivers, T. (2019). Does psychology have a conflict-of-interest problem? *Nature*, *571*(7763), 20.

Collins, D. B., & Holton, E. F. (2004). The effectiveness of managerial leadership

development programs: A meta-analysis of studies from 1982 to 2001. *Human Resource Development Quarterly, 15*, 217-248.

Cooper, W. H., & Richardson, A. J. (1986). Unfair comparisons. *Journal of Applied Psychology*, *71*, 179-184.

Day, D. V. (2001). Leadership development: A review in context. *The Leadership Quarterly, 11*, 581-613.

Day, D., & Dragoni, L. (2015). Leadership Development: Outcome-oriented Review based on Time and Levels of Analyses. *Annual Review of Organizational Psychology and Organizational Behavior*, *2*, 133-156.

Day, D. V., Fleenor, J. W., Atwater, L. E., Sturm, R. E., & McKee, R. A. (2014). Advances in leader and leadership development: A review of 25 years of research and theory. *The Leadership Quarterly*, *25*, 63-82.

Day, D., & Sin, H. P. (2011). Longitudinal tests of an integrative model of leader development: Charting and understanding developmental trajectories. *The Leadership Quarterly*, *22*, 545-560.

DeRue, D.S., Nahrgang, J. D., Hollenbeck, J. R., & Workman, K. (2012). A quasi-experimental study of after-event reviews and leadership development. *Journal of Applied Psychology*, *97*, 997-1015.

Dipietro, R. B. (2006). Return on investment in managerial training: Does the method matter? *Journal of Foodservice Business Research*, *7*, 79-96.

Dvir, T., Eden, D., Avolio, B. J., & Shamir, B. (2002). Impact of transformational leadership on follower development and performance: A field experiment. *Academy of Management Journal*, *45*, 735-744.

Eden, D. (1984). Self-fulfilling prophecy as a management tool: Harnessing Pygmalion. *Academy of Management Review*, *9*, 64-73.

Eden, D. (1992). Leadership and expectations: Pygmalion effects and other self-fulfilling

prophecies in organizations. *The Leadership Quarterly*, *3*, 271-305.

Elo, A. L., Ervasti, J., Kuosma, E., & Mattila-Holappa, P. (2014). Effect of a leadership intervention on subordinate well-being. *Journal of Management Development*, *33*, 182-195.

Feldman, D. C., & Lankau, M. J. (2005). Executive coaching: A review and agenda for future research. *Journal of Management*, *31*, 829-848.

Fischer, T., Dietz, J., & Antonakis, J. (2017). Leadership process models: A review and synthesis. *Journal of Management*, *43*, 1726-1753.

Fitzgerald, S., & Schutte, N.S. (2010). Increasing transformational leadership through enhancing self-efficacy. *Journal of Management Development*, *29*, 495-505.

Gangl, M. (2010). Causal inference in sociological research. *Annual Review of Sociology*, *36*, 21-47.

Gilpin-Jackson, Y., & Bushe, G.R. (2007). Leadership development training transfer: A case study of post-training determinants. *Journal of Management Development*, *26*, 980-1004.

Graen, G., Novak, M.A., & Sommerkamp, P. (1982). The effects of leader-member exchange and job design on productivity and satisfaction: Testing a dual attachment model. *Organizational Behavior and Human Performance*, *30*, 109-131.

Grossman, R., & Salas, E. (2011). The transfer of training: What really matters. *International Journal of Training and Development*, *15*, 103-120.

Heckman, J. J. (2005). The scientific model of causality. *Sociological Methodology*, *35*, 1-97.

Holdnak, B. J., Clemons, T.C., & Bushardt, S. C. (1990). Evaluation of organisation training by the Solomon Four Group design: A field study in self-esteem training. *Journal of Managerial Psychology*, *5*, 25-31.

Hughes, D. J. (2018). Psychometric validity: Establishing the accuracy and appropriateness of psychometric measures. In P. Irwing, T. Booth, & D. J. Hughes (Eds.), *The Wiley handbook of psychometric testing: A multidisciplinary reference on survey, scale and test*

development (pp. 751–779). Chichester, UK: Wiley.

Hughes, D. J., Lee, A., Tian, A. W., Newman, A., & Legood, A. (2018). Leadership, creativity, and innovation: A critical review and practical recommendations. *The Leadership Quarterly*, *29*, 549-569.

Imbens, G. W., & Rubin, D. B. (2017). Rubin causal model. *The new palgrave dictionary of economics*, 1-10.

Ivancevich, J. M. (1982). Subordinates' reactions to performance appraisal interviews: A test of feedback and goal-setting techniques. *Journal of Applied Psychology*, *67*, 581-587.

Jeffreys, H. (1961). *Theory of probability*, 3rd Edition. Oxford: Oxford University Press.

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, *23*, 524-532.

Kirkpatrick, D. L. (1959). Teaching for evaluating training programs. *Journal of American Society of Training Directors*, *13*, 3-9.

Kruschke, J. K (2013). Bayesian estimation supersedes the t test. *Journal of Experimental Psychology: General*, *142*, 573-603

Lacerenza, C. N., Reyes, D. L., Marlow, S. L., Joseph, D. L., & Salas, E. (2017). Leadership training design, delivery, and implantation: A meta-analysis. *Journal of Applied Psychology*, *102*, 1686-1718.

Ladegard, G., & Gjerde, S. (2014). Leadership coaching, leader role-efficacy, and trust in subordinates. A mixed methods study assessing leadership coaching as a leadership development tool. *The Leadership Quarterly*, *25*, 631-646.

Lawler III, E. E. (1977). Adaptive experiments: An approach to organizational behavior research. *Academy of Management Review*, *2*, 576-585.

Leister, A., Borden, D., & Fiedler, F. E. (1977). Validation of contingency model leadership training: Leader Match. *Academy of Management Journal*, *20*, 464-470.

Lonati, S., Quiroga, B. F., Zehnder, C., & Antonakis, J. (2018). On doing relevant and rigorous experiments: Review and recommendations. *Journal of Operations Management*, *64*, 19-40.

Martin, R., Epitropaki, O., & O'Broin, L. (2018). Methodological issues in leadership training research: In pursuit of causality. In *Methodological challenges and advances in managerial and organizational cognition* (pp. 73-94). Emerald Publishing Limited.

McClean, S. T., Barnes, C. M., Courtright, S. H., & Johnson, R. E. (2019). Resetting the clock on dynamic leader behaviors: A conceptual integration and agenda for future research. *Academy of Management Annals*, *13*, 479-508.

McGehee, W., & Gardner, J. E. (1955). Supervisory training and attitude change. *Personnel Psychology*, *8*, 449-460.

Miscenko, D., Guenter, H., & Day, D. V. (2017). Am I a leader? Examining leader identity development over time. *The Leadership Quarterly*, *28*, 605-620.

Moffie, D. J., Calhoon, R., & O'Brien, J. K. (1964). Evaluation of a management development program. *Personnel Psychology*, *17*, 431-440.

Morgan, S. L., & Winship, C. (2015). *Counterfactuals and causal inference* (2nd ed.). New York: Cambridge University Press.

Neck, C. P., & Manz, C.C. (1996). Thought self-leadership: The impact of mental strategies training on employee cognition, behavior, and affect. *Journal of Organizational Behavior*, *17*, 445-467.

Orne, M. T. (1962). On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. *American Psychologist*, *17*, 776-783.

Parry, K. W., & Sinha, P. N. (2005). Researching the trainability of transformational organizational leadership. *Human Resource Development International*, *8*, 165-183.

Paul, W. J., Robertson, K. B., & Herzberg, F. (1969). Job enrichment pays off. *Harvard*

Business Review, *47*, 61-78.

Pearl, J. (2009). Causal inference in statistics: An overview. *Statistics Surveys*, *3*, 96-146.

Podsakoff, P. M., & Podsakoff, N. P. (2019). Experimental designs in management and leadership research: Strengths, limitations, and recommendations for improving publishability. *The Leadership Quarterly*, *30*, 11-33.

Porras, J. J., Hargis, K., Patterson, K. J., Maxfield, D. G., Roberts, N., & Bies, R. J. (1982). Modelling-based organizational development: A longitudinal assessment. *The Journal of Applied Behavioral Science*, *18*, 433-446.

Powell, S. K., & Yalcin, S. (2010). Managerial training effectiveness: A meta-analysis 1952-2002. *Personnel Review*, *39*, 227-241.

Riggio, R. E. (2008). Leadership development: The current state and future expectations. *Consulting Psychology Journal: Practice and Research*, *60*, 383-392.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*, 688-701.

Rubin, D. B. (1975). Bayesian inference for causality: The importance of randomization. In *The Proceedings of the social statistics section of the American Statistical Association* (Vol. 233, p. 239). Alexandria, VA: American Statistical Association.

Russon, C., & Reinelt, C. (2004). The results of an evaluation scan of 55 leadership development programs. *Journal of Leadership & Organizational Studies*, *10*, 104-107.

Rutterford, C., Copas, A., & Eldridge, S. (2015). Methods for sample size determination in cluster randomized trials. *International Journal of Epidemiology*, *44*, 1051-1067.

Shadish, W. R., & Cook, T. D. (2009). The renaissance of field experimentation in evaluating interventions. *Annual Review of Psychology*, *60*, 607-629.

Shaver, J. M. (2019). Causal identification through a cumulative body of research in the study of strategy and organizations. *Journal of Management*,

Smith, R. M., Montello, P. A., & White, P. E. (1992). Investigation of interpersonal management training for educational administrators. *The Journal of Educational Research*, *85*, 242-245.

Solomon, R. L. (1949). An extension of control group design. *Psychological Bulletin, 46*, 137-150.

Steensma, H., & Groeneveld, K. (2010). Evaluating a training using the "four levels model". *Journal of Workplace Learning*, *22*, 319-331.

Stouffer, S. A. (1950). Some observations on study design. *American Journal of Sociology, 55*, 355-361.

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science: A review journal of the Institute of Mathematical Statistics*, *25*, 1-21.

Taylor, P. J., Russ-Eft, D. F., & Taylor, H. (2009). Transfer of management training from alternative perspectives. *Journal of Applied Psychology, 94*, 104-121.

Tharenou, P., & Lyndon, J. T. (1990). The effect of a supervisory development program on leadership style. *Journal of Business and Psychology*, *4*, 365-373.

Van Lent, L. (2007). Endogeneity in management accounting research: A comment. *European Accounting Review, 16*, 197–205.

Wagenmakers, E. J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Love, J., Selker, R., Gronau, Q. F., Šmíra, M., Epskamp, S., Matzke, D., Rouder, J. N., & Morey, R. D. (2018a). Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications. *Psychonomic Bulletin & Review*, *25*, 35-57.

Wagenmakers, E. J., Love, J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Selker, R., Gronau, Q. F., Dropmann, D., Boutin, B.,Meerhoff, F., Knight, P., Raj, A., van Kersteren, K-J., van Doorn, J., Šmíra, M., Epskamp, S., Etz, A., Matzke, D., de Jong, T., van den Bergh, D., Sarafoglou, A., Steingroever, H., Derks, K., Rouder, J. N., & Morey, R. D. (2018b). Bayesian

inference for psychology. Part II: Example applications with JASP. *Psychonomic Bulletin & Review*, *25*, 58-76.

West, S. G. (2009). Alternatives to randomized experiments. *Current Directions in Psychological Science*, *18*, 299-304.

Yeow, J., & Martin, R. (2013). The role of self-regulation in developing leaders: A longitudinal field experiment. *The Leadership Quarterly*, *24*, 625-637.

Table 1: Internal Threats to Validity in Leadership Development Research (adapted from Martin, Epitropaki, & O'Broin, 2018)

INTERNAL THREATS		
Threat	**Description**	**Example Application to Leadership Training**
History	Events, other than the experimental conditions, influence the results.	Changes in environment, other than the training, affect changes between pre-test and post-test (such as changes to product, processes, market conditions, and working conditions).
Maturation	During the study, changes not due to the treatments within the participants affect outcomes.	Developmental improvements within the leaders and/or team members over time that are independent of the training.
Testing	Exposure to a pre-test or intervening assessment influences performance on a post-test.	Pre-test measures (e.g. leader's relationship skills) make salient factors that the followers had hitherto not considered which renders them more sensitive to these factors at post-test (e.g., if manager shows improved relationship skills).
Instrumentation	Testing instruments are not consistent; or pre-test and post-test are not equivalent.	Different types of measures, measurement techniques or order of evaluation are employed pre-test and post-test to capture similar constructs.
Statistical Regression	Scores of participants are initially high or low and regress towards the mean during retesting.	If leaders initially have very high scores on the leadership capabilities in the training, then these scores are unable to improve and can decrease giving false estimates of change.
Selection	Systematic differences exist in participants' characteristics between treatment groups.	If allocation of leaders between training and non-training conditions is not random, then pre-existing differences between the conditions might explain difference in post-test outcomes.

Experimental Mortality	Participants' attrition during the study may bias results.	Some leaders (e.g., poor performers) leave the organization during the evaluation period such that the sample populations differ between conditions and over time.
Selection-maturation Interaction	The selection of comparison groups and maturation interact so that some groups change differently to others.	Team members in a leader non-training condition maturate (i.e., change naturally over time) at a different rate than those in the leader training condition.

Table 2: The ability of different research designs to deal with threats to internal validity

Research Design	Design	History	Matura-tion	Testing	Instru-mentation	Regress-ion	Selection	Mortality	Selection x Maturation
Intervention Only									
One Shot	X – O	-	-				-	-	
Repeated Measures	O – X – O	-	-	-	-	?	+	+	-
Time series	O – O – X – O – O	-	+	+	?	+	+	+	+
Independent Groups									
Random/Matched Allocation (Laboratory/Field Experiments)	X – O C – O	+	+	+	+	+	+	+	+
Non-random Allocation (Quasi-experiments)	X – O C – O	+	?	+	+	+	-	-	-
Independent Groups with Repeated Measures									
Random/Matched Allocation (Laboratory/Field Experiments)	O – X – O O – C – O	+	+	+	+	+	+	+	+
Non-random Allocation (Quasi-experiments)	O – X – O O – C – O	+	+	+	+	+	-	+	-
Solomon Design	O – X – O O – C – O X – O O	+	+	+	+	+	+	+	+

Table 3: Summary of codings for leadership training studies

	Sample Representation	Condition Randomization	Condition Independence	Temporal Deisgn	Similar time Measurement	Pre-test Differences	Author(s) Involvement
Yes	18	28	10	1	14	10	34
No	44	34	44	61	28	13	15
Not coded	0	0	8	0	20	9	13

N = 62, except pre-test differences which was 32